ECE260B – CSE241A
Winter 2004

On-Chip Memory

Website: http://vlsicad.ucsd.edu/courses/ece260b-w04
Outline

- Embedded Memories today
  - References: Rabaey, Veendrick

- Volatile Memories today
  - FIFO
  - eRAM

- Non-Volatile Memories today

- Memories in the future
  - Magneto-resistive RAM (MRAM)
  - Ferro-electric RAM (FeRAM/ FRAM)
  - Ovonics Unified Memory (OUM)
Semiconductor Memory Trends

Today, about 1-2Gbit DRAM
SRAM: 512Mbit available as separate products, 2MB on-chip cache in servers

Memory size as function of time: x4 every 3-4 years
Embedded Memories Today

- Embedded memory on every chip
  - FIFO – for buffering
  - Registers – for pipelining and temporary data storage
  - RAM – temporary data storage (for high throughput)
  - ROM – storing standard instruction set in micro-controllers, power-on sequence

- More than 70% of chip area is memory
  - high density, good yield - very important
Outline

- Embedded Memories today
- Volatile Memories today
  - FIFO / Register
  - eRAM
- Non-Volatile Memories today
- Memories in the future
  - Magneto-resistive RAM (MRAM)
  - Ferro-electric RAM (FeRAM/ FRAM)
  - Ovonics Unified Memory (OUM)
FIFO / Register files

- 7-T / 8-T memory cell
- Separate Read/Write bitlines to counter metastability
- Counters in FIFO
- Multiple ports in register files. Separate lines for “0” and “1” read
RAM Architecture

- Address Decoders
  - X and Y-decoder
- Memory Cell Array
- Sense Amplifier
- Write Circuit
- I/O buffer
- Mostly Self-Timed
Generic Memory Structure

- Row Address Decoder
  - $A_1$, $A_2$, $A_3$, ..., $A_N$

- Column Decoders
  - $B_1$, $B_2$, $B_3$, $B_4$, ..., $B_M$

- Bit-line
- Word-line
- Row 1, Row 2, Row $2^N$

- Memory Cell
  - $(2^N \times 2^M$ total)

- Number of bits $= 2^{M+N}$

- Selects one or more bit lines
Memory Architecture: Decoders

N words => N select signals
Too many select signals

Decoder reduces # of select signals
\[ K = \log_2 N \]

Source: D. Sylvester, U. Michigan
Array-Structured Memory Architecture

Problem: ASPECT RATIO or HEIGHT >> WIDTH

- Amplify swing to rail-to-rail amplitude
- Selects appropriate word

Source: D. Sylvester, U. Michigan
Hierarchical Memory Architecture

Advantages:
1. Shorter wires within blocks
2. Block address activates only 1 block => power savings

Source: D. Sylvester, U. Michigan
Memory Timing: Definitions

Read Cycle

Read Access

Data Valid

Write Access

Write Cycle

Data Written

Source: D. Sylvester, U. Michigan
Read-Write Memories (RAM)

- **STATIC (SRAM)**
  
  Data stored as long as supply is applied
  Large (6 transistors/cell)
  Fast
  Differential

- **DYNAMIC (DRAM)**
  
  Periodic refresh required
  Small (1-3 transistors/cell)
  Slower
  Single Ended

Source: D. Sylvester, U. Michigan
6-Transistor CMOS SRAM Cell

Source: D. Sylvester, U. Michigan
6T-SRAM — Layout

- Extremely dense
- Modern processes can fit a 6T SRAM cell in \( \sim 1.3\mu m^2 \)

Source: D. Sylvester, U. Michigan
1-Transistor DRAM Cell

Write: $C_S$ is charged or discharged by asserting $WL$ and $BL$.
Read: Charge redistribution takes place between bit line and storage capacitance.

$$\Delta V = V_{BL} - V_{PRE} = (V_{BIT} - V_{PRE}) \frac{C_S}{C_S + C_{BL}}$$

Voltage swing is small; typically around 250 mV.

Source: D. Sylvester, U. Michigan
DRAM Cell Observations

- 1T DRAM requires sense amplifier for each bit line, due to charge redistribution read-out.
- DRAM memory cells are single ended in contrast to SRAM cells.
- Read-out of 1T DRAM cell is destructive; read and refresh operations are necessary for correct operation.
- 1T cell requires presence of an extra capacitance that must be explicitly included in the design.

Source: D. Sylvester, U. Michigan
1-T DRAM Cell

(a) Cross-section

- Metal word line
- $n^+$ diffusion
- Polysilicon
- Inversion layer induced by plate bias
- *poly*

(b) Layout

- Capacitor
- M1 word line
- Polysilicon gate
- Polysilicon plate
- Diffused bit line
- *poly*

Uses Polysilicon-Diffusion Capacitance

Expensive in Area

Source: D. Sylvester, U. Michigan
Advanced 1T DRAM Cells

- Trench Cell
  - Cell Plate Si
  - Capacitor Insulator
  - Storage Node Poly
  - 2nd Field Oxide
  - Si Substrate

- Stacked-capacitor Cell
  - Word line
  - Insulating Layer
  - Cell plate
  - Capacitor dielectric layer
  - Transfer gate
  - Isolation
  - Storage electrode

Source: D. Sylvester, U. Michigan
Embedded RAM

- SRAMs and DRAMs

<table>
<thead>
<tr>
<th>SRAM</th>
<th>DRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-T / 4-T memory cell</td>
<td>Capacitor based storage. High Density</td>
</tr>
<tr>
<td>Low Power – important requirement for system on chip</td>
<td>Refresh cycles required – hence high power</td>
</tr>
<tr>
<td>Fast Data Access</td>
<td>Slower Access cycles</td>
</tr>
<tr>
<td>Relative transistor sizes determine Noise Margin</td>
<td>Capacitor size determines Noise Margin</td>
</tr>
</tbody>
</table>

- Noise Margin
  - Important figure of merit
  - Degraded with scaling
Embedded RAM

- Integrating DRAMs with CMOS
  - Very high density achievable
  - Promises good on-chip speed
  - Extra mask step even for Logic region of chip
  - Large on-chip power
  - However, trend is towards integration

Capacitor consists of:
- MIS – metal–insulator–semiconductor
- MIM – metal-insulator-metal

source nec.com
Outline

- Embedded Memories today
- Volatile Memories today
  - FIFO
  - e-RAM
- Non-Volatile Memories today
- Memories in the future
  - Magneto-resistive RAM (MRAM)
  - Ferro-electric RAM (FeRAM/ FRAM)
  - Ovonics Unified Memory (OUM)
Non-Volatile Memories

- **Standard ROM**
  - Programmed during fabrication
  - Diffusion programmable / metal or via programmable options

- **One Time Programmable (OTP) ROM**
  - Involves blowing of fuses – after fabrication

- **Erasable Programmable ROM (EPROM)**
  - Erase and Program through UV light application

- **Electrically Erasable Programmable ROM (EEPROM)**
  - Programmable by application of high voltage
  - Involves two supply voltages – normally not a problem for today’s chips
ROM

Figure 10.8  A 4 × 4 bipolar ROM cell array.
### ROM Cell

- **Standard ROM and OTPROM** involve setting of switch
- **Erasable ROM** involve floating gate
  - Charge stored on floating gate determines data stored
- **Floating gate memories** can be multi-level in nature (providing still higher density)

![ROM cell](diffusion/contact/metal programming)

![EPROM cell](FG programmed by UV)
Floating Gate (Programming)

Figure 10.20  Programming the floating-gate transistor.
EEPROM & Flash

- Apply Fowler-Nordheim tunneling to store charge in gate oxide

- Flash has burst mode read/write (writes/erases 512 bytes (or more) in one cycle – hence faster

- Require multiple voltage supplies on chip
  - High Voltage used for erase and program operations (order 10V presently)
  - Standard (Low) voltage used for read operation – Lower operational power

Threshold voltage control by charge on floating gate
Flash

- Two types – NOR and NAND
  - NOR type more widely used because of lower supply voltage requirement
  - NAND requires (order 20V supply for Write/Erase)

- Multiple Level Programming
  - Increased density (2 or more bits per cell)
  - Increased complexity – still more supply voltages
  - Advanced current sense circuits

- One crude embeddable solution by Virage Logic

Threshold voltage on chip for single bit and 2-bit Flash memory cells

source Intel.com
Outline

- Embedded Memories today
- Volatile Memories today
  - FIFO
  - e-RAM
- Non-Volatile Memories today
  - ROM
  - Flash
- Memories in the future
  - Magneto-resistive RAM (MRAM)
  - Ferro-electric RAM (FeRAM/ FRAM)
  - Ovonics Unified Memory (OUM)
Memories tomorrow #1

- Requirement for embedded non-volatile memory
  - Numerous applications like PDAs, Digital Cameras

- Less Area
  - Enable huge data-storage on chip

- Should have fast access
  - Unlike FLASH

- Unlimited Read / Write
  - Important requirement for SoC

- Completely Electrical Operation
Memories tomorrow #2

- **Magneto-resistive RAM (~2004)**
  - IBM, Motorola, Infineon, Nonvolatile Electronics (NVE)

- **Ferro-electric RAM (FRAM/ FeRAM) (~ 2004)**
  - Ramtron, Symetrix, Fujitsu, Toshiba, IBM/ Infineon, Samsung, Motorola, Hitachi, Matsuhita, Micron

- **Ovonics Unified Memory (OUM) (~2004)**
  - Ovonyx, Intel, STMicroelectronics, British Aerospace

- **Nano-Floating Gate memory (~2005)**

- **Single/ Few electron memories (SET) (~2007)**

- **Molecular memories (~2010)**
Outline

- Embedded Memories today
- Volatile Memories today
  - FIFO
  - e-RAM
- Non-Volatile Memories today
  - ROM
  - Flash
- Memories in the future
  - Magneto-resistive RAM (MRAM)
  - Ferro-electric RAM (FeRAM/ FRAM)
  - Ovonics Unified Memory (OUM)
Magneto-resistive RAM #1

- Uses thin magnetic films to store data on hysteresis
  - Based on Giant Magneto-resistance (GMR)

- 1MTJ-1T memory cell

- Requires magnetic tunnel junction material
  - Stack of thin MTJ layers separated by dielectric
  - Data stored in direction of magnetic moment of free MTJ layer
  - Change in polarity measured as resistance change

- Cross point matrix architecture

Source: Durlam et al. ISSCC 2000
Magneto-resistive RAM #2

- Read Operation requires switching ON of isolation (pass) transistor
- Reference BL required for sensing resistance (hence data)

- Write Operation involves flow of program current while isolation transistor is OFF
  - Only one cell has 2 perpendicular program currents to enable write

Source: Durlam et al., ISSCC 2000
Magneto-resistive RAM # 3

- High Density
- Non Destructive Read
- Low Voltage and Low Power
- Unlimited R/W endurance
- Material compatibility with CMOS process is a key challenge
Outline

- Embedded Memories today
- Volatile Memories today
  - FIFO
  - e-RAM
- Non-Volatile Memories today
  - ROM
  - Flash
- Memories in the future
  - Magneto-resistive RAM (MRAM)
  - Ferro-electric RAM (FeRAM/ FRAM)
  - Ovonics Unified Memory (OUM)
Ferro-electric RAM #1

- Uses bistable crystalline materials
- 1T-1C cell
- Data stored by applying voltage to polarize internal dipoles “Up” or “Down”
- No Iron – no magnetism
- Ferro-electric capacitor
  - Has dielectric made of ferro-electric material
  - Lead Zirconate Titanate (PZT) is the most widely used family Pb(Zr$_x$Ti$_{1-x}$)O$_3$

Data “1” and “0” in dielectric
Ferro-electric RAM #2

- Read operation similar to SRAM
- Boosted WL required for Write
- Generation of reference voltage required for sensing stored data

Source: Jones, CICC 1998
Ferro-electric RAM #3

- “Fast” Random Read Access
- Fast Write with very low power consumption
- Destructive Read
- Limited read and write cycles (not in absolute terms, but since read is destructive, write mandatory after/with read)
Outline

- Embedded Memories today
- Volatile Memories today
  - FIFO
  - e-RAM
- Non-Volatile Memories today
  - ROM
  - Flash
- Memories in the future
  - Magneto-resistive RAM (MRAM)
  - Ferro-electric RAM (FeRAM/ FRAM)
  - Ovonics Unified Memory (OUM)
Ovonics Unified Memory #1

- Uses Chalcogenide material alloys used in re-writeable CDs and DVDs (Ge, Sb, Te alloy)
- 1T-1C cell
- Electric Energy (heat) converts material between crystalline (conductive) and amorphous (resistive) phases
- Data read by measuring resistance

Source: Gill et al.
Ovonics Unified Memory #2

- Read Operation exactly same as standard SRAM
- Age independent 40X change in resistance in “0” and “1” states
- “Write 0” involves a high peak pulse for a short duration of time
- “Write 1” (low resistive state) involves a lower current pulse for a longer duration of time

Source: Gill et al.
Ovonics Unified Memory #3

- High Density
- Non-destructive Read
- Low voltage
- Low Power
- \(\sim 10^{12}\) write/ erase cycles
- Easy to integrate with logic
Outline

- Embedded Memories today

- Volatile Memories today
  - FIFO
  - e-RAM

- Non-Volatile Memories today
  - ROM
  - Flash

- Memories in the future
  - Magneto-resistive RAM (MRAM)
  - Ferro-electric RAM (FeRAM/ FRAM)
  - Ovonics Unified Memory (OUM)
## Embedded Memory Landscape

<table>
<thead>
<tr>
<th>Storage Mechanism</th>
<th>DRAM</th>
<th>SRAM</th>
<th>Flash</th>
<th>MRAM</th>
<th>FRAM</th>
<th>OUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Feature size</td>
<td>130nm</td>
<td>130nm</td>
<td>150nm</td>
<td>350nm</td>
<td>130nm</td>
<td>100nm</td>
</tr>
<tr>
<td>Cell Structure</td>
<td>1T-1C</td>
<td>6T</td>
<td>1T</td>
<td>1T-1C</td>
<td>1T-1C</td>
<td>1T-1C</td>
</tr>
<tr>
<td>Cell Size (area)</td>
<td>8F²</td>
<td>10F²</td>
<td>10F²</td>
<td>~40F²</td>
<td>20-40F²</td>
<td>6F²</td>
</tr>
<tr>
<td></td>
<td>0.14 μ²</td>
<td>0.16 μ²</td>
<td>0.19 μ²</td>
<td>4.9 μ²</td>
<td>0.68 μ²</td>
<td>0.06 μ²</td>
</tr>
<tr>
<td>R/W time</td>
<td>&lt;20ns</td>
<td>&lt;10ns</td>
<td>~80ns (R)</td>
<td>&lt;25ns</td>
<td>&lt;10ns</td>
<td>&lt;100ns</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1ms (W)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of cycles</td>
<td>Unlimited</td>
<td>Unlimited</td>
<td>&gt;1E5</td>
<td>&gt;1E15</td>
<td>&gt;1E13</td>
<td>&gt;1E13</td>
</tr>
<tr>
<td>Retention</td>
<td>Volatile</td>
<td>Volatile</td>
<td>&gt;10yrs</td>
<td>&gt;10yrs</td>
<td>&gt;10yrs</td>
<td>&gt;10yrs</td>
</tr>
</tbody>
</table>
Embedded Memory Landscape (optional)

![Diagram showing the embedded memory landscape with different memory types and their characteristics.](Image)

- New NV RAM Space: FeRAM, OUM, MRAM
  - Unlimited Read Cycles
- Volatile Memory Space: SRAM, DRAM
- Flash Space: ETOX NAND
  - Unlimited Read Cycles
- ROM Space: ROM, EPROM
  - Not in-system changeable
  - Unlimited Read Cycles

Source: Gill et al.
Content Addressable Memories (CAMs)

- A different league of memories

- Provide output by checking if content is stored
  - Reverse functionality

- Can be used in
  - Implementing redundancy in memories (check address bus and re-route read/write request)
  - Routers and other network components to route data to particular location
    - Promise immense gain in speed
    - Bottleneck - efficiency of data storage
**CAMs (#2)**

- **Match operation**
  - compare data on D bus with data stored in memory cell (WL is low)

- **When full data matched** – MATCH remains high

- **Matching can be full (all bits) or partial (some bits) dependent on architecture**

- **Ternary CAMs also in market**
  - Useful for high level data routing
Ternary CAMs

- Ternary architecture generally used in routers and network processors

16-T ternary SRAM based CAM cell

8-T DRAM based ternary CAM cell
DRAMs

- **Standard DRAM**
  - Row Address Select (RAS) and Column Address Select (CAS) inputs determine data change

- **SDR-DRAM**
  - RAS and CAS signals activated on rising edge of the clock - synchronous

- **DDR (Double Data Rate) DRAM**
  - Address latched at both rising and falling Clock edges – double data rate

Source – skeeterbytes.com
Resistive-load SRAM Cell

Static power dissipation – must keep $R_L$ big to limit this

Large $R_L = $ large delay pulling up $\rightarrow$ precharge bit lines high so read out only pulls low

Source: D. Sylvester, U. Michigan

Andrew B. Kahng, UCSD